Estimation of mitochondrial DNA copy number from high-throughput microarray data

Overview

Mitochondria are organelles that are found in most human cells and that present their own specific DNA (mitochondrial DNA), which is separate from the cell's nuclear DNA. The amount of mitochondrial DNA in a cell has recently been identified as a key biomarker for cardiovascular and neurodegenerative diseases. Currently, quantitative polymerase chain reaction (qPCR) is the state of the art method for measuring mitochondrial DNA contents in human cells. However, qPCR is scarcely used in practice or in studies because of its cost.

In the past three decades, two other methods have been used to sequence mitochondrial DNA: Whole Genome Sequencing (WGS), and high-throughput DNA microarrays. Although informative, these methods do not offer a clear way of estimating the absolute amount of mitochondrial DNA per cell in a sample. Much work has already been done for the case of WGS data. However, as of today, an accurate way to approximate the amount of mitochondrial DNA in a sample using high-throughput DNA microarrays does not exist.

In this report, we hypothesise that the amount of mitochondrial DNA in a sample can be accurately estimated using microarray data. To do so, we start by preprocessing the data to remove as many as possible potential biases, and develop a novel binning technique to overcome the difficulty of the wide variety of microarrays used in studies. Then, we estimate the amount of mitochondrial DNA in cells using a neural network whose inputs contain both binned probe signal intensities of the mitochondrial genome on a standard high-throughput DNA micro-array, and principal components from a principal components analysis of the autosomal DNA data.

At the time of writing, although they are too weak to call for reliable and accurate estimations, our method achieves more accurate and more consistent results than the current techniques for estimating mitochondrial DNA contents in a sample using high-throughput DNA microarrays. As standard metrics like MSE, RMSE and MAE make less sense when compared across different studies with different DNA microarrays, we use the R^2 of a standard OLS regression of the estimates on the truth value and the Pearson correlation coefficient to evaluate our results against the literature on the subject. We achieve a $R^2 = 0.29$ on the test dataset and a Pearson correlation coefficient of 54%.

Table of contents

1 Introduction

1.1 Current methods for mtDNA copy number estimation

Mitochondrial copy number (a proxy measure for the mitochondrial DNA contents within a cell) has recently been identified as a potential key biomarker for the overall metabolism of mitochondria and multiple phenotypes and diseases, including cardiovascular and neurodegenerative ones. Currently, quantitative polymerase chain reaction (qPCR) is the state-of-the-art method for measuring mitochondrial DNA content and assessing its variations between individuals. qPCR is nevertheless scarcely used, mainly as a result of both its cost, availability and the difficulty to make it reproducible for studies with thousands of samples.

Genetic data gathering can, however, be achieved without qPCR. There mostly are two main supplementary methods that allow inferring mitochondrial DNA sequencing data: Whole-Genome sequencing (WGS), and high-throughput DNA microarrays. In the past few years, both of these methods have been investigated for mitochondrial copy number estimation. While WGS remains costly and, to a lesser extent than qPCR, hardly reproducible in broad studies, the cost per patient for high-throughput DNA microarrays has continuously decreased.

Large studies have already been led using this method (1000 Genomes, UK Biobank) and most of them include mitochondrial DNA sequencing data. An efficient way to approximate mitochondrial DNA copy number using array intensity data from high-throughput DNA microarrays does however not exist at this time.

We hypothesise that variations in mitochondrial DNA content between individuals can be accurately and systematically estimated using high-throughput DNA microarray data and that it can be done in a way that does not only work on a single type of array, but that can be generalised over different types of arrays containing different, potentially non-overlapping mitochondrial intensity probes.

1.2 Contents and outline

In the following report, our main aim is to show that mitochondrial copy number can be approximated with a simple neural network whose inputs are binned probe signal intensities of the mitochondrial genome on a standard high-throughput DNA microarray, and additional features gathered directly from autosomal DNA whose aim are to correct for potential unmeasured confounding.

Although too weak to call for reliable estimations yet, our method achieves more accurate and more consistent results than the current state-of-the-art in mitochondrial copy number estimation using high-throughput microarrays. After going over the extent and definitions of the issue as well as the data in the first section, we develop our estimation procedure in the second section, and evaluate it using standard metrics from the literature on the subject and known covariates in the third section.

2 An overview of mitochondrial DNA contents

A mitochondrion (plural. mitochondria) is an organelle (a specialised subunit of a cell) present in the vast majority of organisms. The most widely accepted hypothesis concerning its origin is the endosymbiotic theory, which dates the presence of mitochondria in eukaryotic cells to billions of years ago, and which claims that mitochondria come from the incorporation (or endocytosis) of a bacteria (α-proteobacterium) into a host cell. A key point of the endosymbiotic theory relies on the fact that mitochondria have their own DNA, which implies that their origin is exogenous. As a matter of fact, the DNA of mitochondria is different from that of the nucleus and is generally transmitted by the mother.

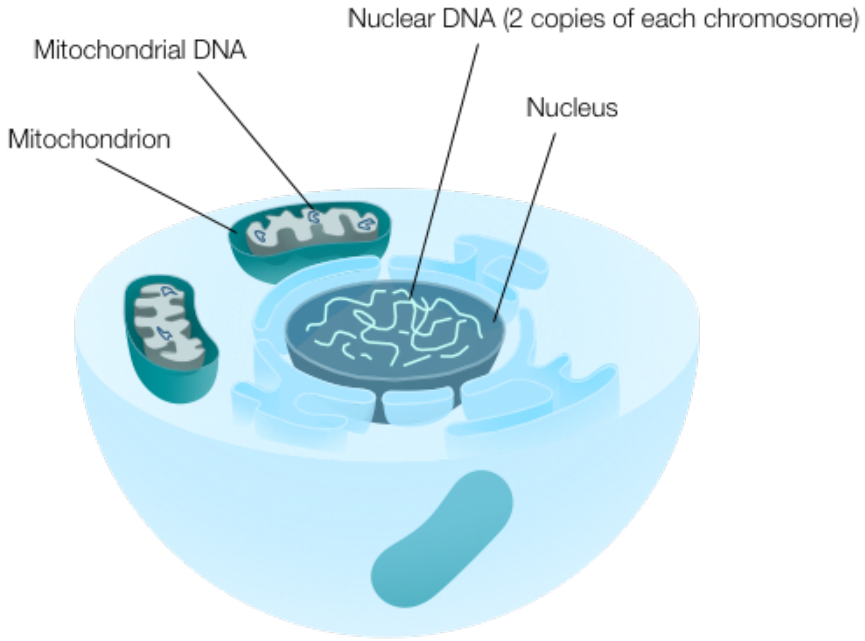

Figure 1: A simplified human cell with the nucleus and mitochondria

Most cells within the human body contain mitochondria, although there are a few types which do not contain any, red blood cells being the most common cell type of the sort. A mitochondrion is usually surrounded by a double membrane, each composed of a phospholipid double layer. Its diameter generally varies between 0.75 and 3 micrometres, while although they are commonly represented as a cylinder for simplification, their general shape and structure are extremely variable. Mitochondria have furthermore been shown to be particularly dynamic organelles that constantly fuse and divide in complex processes that are important for the maintenance of mitochondrial functions [Westermann, 2010].

2.1 The powerhouse of the cell

Mitochondria are involved in a very diverse set of functions within the cells of eukaryotic organisms. Although their functions do not limit themselves to this, their most

widely described use is that of the so-called *powerhouse of the cell*.

The double membrane bound organelles are typically known for producing the majority of the cells supply of adenosine triphosphate. Adenosine triphosphate, or ATP, is a nucleotide which, in all living organisms, provides the energy necessary for chemical reactions, including metabolism, locomotion, cell division, or active transport of chemical species across biological membranes. ATP is thus known as the molecular unit of currency and is necessary for staying alive.

Although there are other ways through which the human body produces the ATP it uses for chemical reactions, the vast majority (around 80%) comes from mitochondria through an enzymatic process known as ATP synthase, that operates between the inner membrane and the matrix [Stock et al., 1999], thus the name of *powerhouse*. Mitochondria are therefore preferentially placed near ATP-consuming cellular areas.

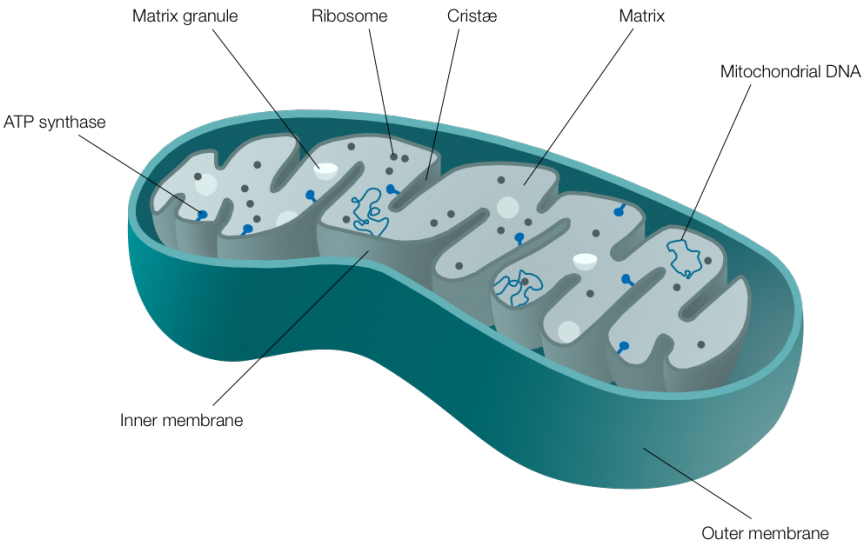

Figure 2: A simple representation of a mitochondrion

Apart from their role in energy production, mitochondria have also widely been investigated for their use in apoptosis, cellular differentiation and cell growth [Nunnari and Suomalainen, 2012]. But they were also recently shown to be involved in more diverse and less studied functions like cell signalling and overall regulation of the cellular metabolism [McBride et al., 2006].

Thus, mitochondria play a key role in health and, unsurprisingly, mitochondrial dysfunctions have been shown to take part in many diseases, including neurodegenerative disorders [Cho et al., 2010; Chan et al., 2011], cardiovascular diseases [Götz et al., 2011], cancer [Scatena, 2012], obesity [Pietiläinen et al., 2008] and type I and type II diabetes [Lowell and Shulman, 2005].

2.2 The mitochondrial DNA copy number

As the endosymbiotic theory suggests, mitochondria have their independent genome that shows substantial similarity to bacterial genomes [Andersson et al., 2003]. Although similar to that of bacteria, the mitochondrial genome today contains only a very small part of its original code. The sequences eliminated from the mitochondrial genome have either been transferred to the nucleus or lost.

In humans, however, although the mitochondrial genome and the nuclear one interact in numerous ways, the mitochondrial genome differs a lot from the nuclear genome. Despite it being still debated, it is widely believed that, unlike the nuclear genome which is inherited from both the mother and the father, the mitochondrial DNA is entirely matrilineally inherited.

The structure of the mitochondrial genome also differs from that of the nuclear one by its circularity. It consists of 16,569 base pairs that can be split in 37 genes which code for 13 proteins, 22 transfer RNAs and 2 ribosomal RNAs. The genes of the mitochondrial genome are arranged one after the other and are separated only by short non-coding regions (in white on Figure 3). The genes coding for proteins are separated from each other by genes coding for transfer RNAs. A 600bp region called the displacement loop (or D-loop) contains the genomic code for the initiation of the transcription of the genome.

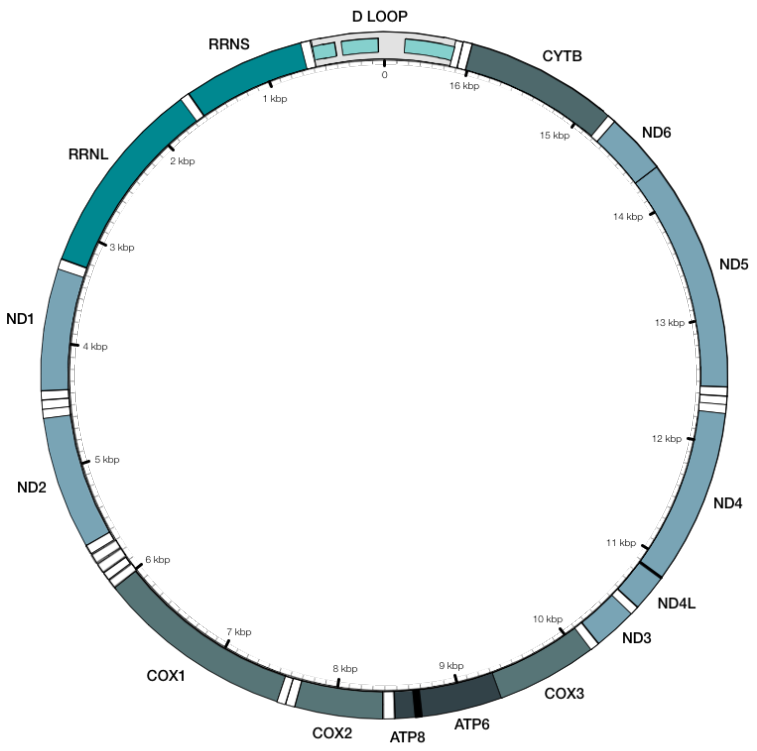

Figure 3: The human mitochondrial circular DNA

Mitochondria also vary in their genomic aspects from the nucleus by the number of copies of the genome that they usually hold. Unlike nuclei, which in humans usually contain two copies of each chromosome, mitochondria can generally contain between 2 and 10 copies of their circular 16,569-base-pairs-long DNA per organelle. Each cell can furthermore have between 10 and 10,000 organelles, depending on the cell's type and functions. For instance, larger cells that are more prone to energy-driven tasks, like neurones, usually contain a high number of mitochondria. While nuclear DNA can therefore usually be found in two copies in most cells of the human body, the amount of mitochondrial DNA content within a cell (the mitochondrial DNA copy number) varies a lot, typically in a range between 20 and 20000 copies.

2.3 Insights on mtDNA as a potential biomarker

In recent years, the amount of mitochondrial DNA contained within specific cells (measured by the mitochondrial DNA copy number, abbreviated as mtDNA CN) was shown to be highly correlated to, and sometimes a good predictor of a diversity of conditions including cancer, obesity, neurodegenerative and cardiovascular diseases and organ failures.

Although debated, most of the literature on the subject argues that dysfunctions in mitochondria could be either the cause, consequence, or confounder of those conditions and that these same dysfunctions result in a decrease or increase in mitochondrial DNA content per cell. As of today, various measures of mitochondrial DNA copy number have been proposed as potentially efficient biomarkers in the screening of multiple diseases [Malik and Czajka, 2013].

In recent works, it was for instance shown that there was a significant reduction in mitochondrial DNA copy number in patients with Parkinson's disease, both through the identification of abnormal quantities of freely circulating mitochondrial DNA in blood [Pyle et al., 2015] and through the quantification of mtDNA content in peripheral blood samples, in substantia nigra pars compacta tissue samples and in frontal cortex tissue samples [Pyle et al., 2016]. Patients with Alzheimer disease have also displayed freely circulating mitochondrial DNA in cerebrospinal fluid [Podlesniy et al., 2013], as well as lower mitochondrial DNA copy numbers in both brain tissue samples [Wei et al., 2017] and blood cells [Delbarba et al., 2016]. A reduction in mitochondrial DNA copy numbers was also observed in patients with Type II Diabetes [Wong et al., 2009] and gestational Type II Diabetes [Weijers and Bekedam, 2007]. Reduced mitochondrial DNA content was also shown to be associated with mortality and infections in patients with chronic kidney disease [Fazzini et al., 2019]. Regarding tumours finally, some cancers, especially of the bladder, breast, and kidney, were shown to be depleted of mitochondrial DNA in comparison to normal tissues [Reznik et al., 2016].

In the other hand, an increase in mitochondrial DNA copy number in urinary samples was shown to be a predictive biomarker of renal injury and function in humans with acute kidney failure (AKI) following cardiac surgery [Whitaker et al., 2015]. Patients with severe sepsis [peng Yan et al., 2018] were also shown to have higher than usual plasma mitochondrial DNA levels.

Concerning cardiovascular diseases, patients who have suffered from acute myocardial infarction were shown to have significantly higher nuclear and mitochondrial DNA levels, suggesting that the mitochondrial DNA copy number could potentially be a novel biomarkers of the disease.[Wang et al., 2015]. Mitochondrial DNA copy number was also independently associated with cardiovascular diseases of different sorts, heart disease and strokes, and was therefore suggested to have clinical utility in improving cardiovascular diseases risk classification. [Ashar et al., 2017]. Finally, mitochondrial DNA copy numbers in peripheral blood samples were shown to be inversely associated with the risk of sudden cardiac arrest [Zhang et al., 2017b].

2.4 Nuclear and mtDNA sequencing methods

The literature on mitochondrial copy number has recently seen great improvements in measurement techniques. At present, there exist various procedures for estimating the mitochondrial copy number of a sample, most of which rely on approximating the ratio between mitochondrial and nuclear genome contents to establish a viable measure to compare individuals. However, these methods depend critically on the type of data available and vary a lot in terms of cost and accuracy. There is currently no systematic way to estimate mitochondrial copy numbers quickly, efficiently and accurately for very large samples at a relatively low expense. Current methods include quantitative real-time polymerase chain reaction (qPCR), Whole-Genome sequencing (WGS) and high-throughput DNA microarrays.

2.4.1 Quantitative Polymerase Chain Reaction (qPCR)

Quantitative Polymerase Chain Reaction (qPCR) was shown to be the current most stable and reliable method in estimating mitochondrial DNA copy number variations between individuals [Malik and Czajka, 2013], and is for these reasons considered by the academic literature as the current gold-standard in terms of measurement.

qPCR works by monitoring in real-time the amplification of DNA molecules during a polymerase chain reaction, that is a thermal cycling method used to rapidly make copies of a DNA sample. The process of estimating mitochondrial DNA copy number by qPCR is, however, costly and time-consuming and can as such not be used for studies with multiple thousands of samples [Fazzini et al., 2018], but is rather often used for smaller studies or for evaluating the efficiency and precision of other methods.

2.4.2 Whole-Genome sequencing (WGS)

Mitochondrial copy number per cell can also be accurately inferred from DNA sequencing read counts using Whole-Genome (or Whole-Exome) Sequencing data [Ding et al., 2015; Chu et al., 2012; Zhang et al., 2017a]. The imputation of the mitochondrial DNA copy number relies on the idea that the average DNA sequencing coverage is proportional to the actual DNA copy number for both nuclear and mitochondrial DNA, thus allowing a weighted ratio between mitochondrial and autosomal coverages (1) to be a good estimate of the mitochondrial DNA copy number. In simpler terms, the copy number can be estimated by the ratio between the number of mitochondrial

DNA sequences in the data for one individual and the number of autosomal DNA sequences for the same individual, thus measuring mitochondrial DNA proportionally to the overall amount of DNA inferred from the Whole-Genome sequencing procedure. The copy number can therefore be estimated using a simple formula (Equation 1).

$$\text{mtDNA copy number} = \frac{\text{mtDNA average coverage}}{\text{autosomal DNA average coverage}} \times 2 \tag{1}$$

Whole-Genome sequencing datasets are usually very large (more than 20 gigabytes per individual on average), and while the task is seemingly simple, identifying the sequences which belong to the genome of autosomal chromosomes to those of the mitochondrial DNA and calculating the average coverage for multiple chromosomes can in practice take a lot of time and resources for studies with thousands of samples (multiple terabytes of data). To reduce the computing time, it was subsequently shown that using randomly selected portions of the nuclear genome to calculate the average autosomal DNA coverage provided accurate results as well [Qian et al., 2017].

In terms of accuracy and consistency, estimates from Whole-Genome sequencing have been shown to be significantly associated to certain medical conditions [Tin et al., 2016; Cai et al., 2015], to all known mitochondrial DNA copy number covariates (age, sex, white blood cell count, etc.) [Longchamps et al., 2019] and to be highly correlated to qPCR estimates [Ding et al., 2015]. Significant and recurrent biases have however also been identified in their outputs [van Heesch et al., 2013].

2.4.3 High-throughput DNA microarrays

Finally, recent works have shown that, to a certain extent, some information on mitochondrial DNA copy numbers could also be derived from high-throughput DNA microarray data [Ding et al., 2015], and that the inferred estimates were associated to all known mitochondrial DNA copy number covariates, as well as to qPCR and WGS estimates (`http://genvisis.org/MitoPipeline/`).

DNA microarrays usually quantify probe-target hybridisation by detection of fluorophore. The principle at the heart of high-throughput DNA microarrays relies on exposing chromosomes (for instance from blood samples) to small known DNA sequences called probes, that have fluorescent molecules attached to them which reveal the presence of the sequence in the sample DNA. By knowing the position of each probe, its sequence and the fluorescence level given by the probe-target hybridisation, it is possible to estimate the probability that each sequence belongs to the DNA of the given sample and, in other words, to capture information on its genome.

Studies conducted using these novel estimates have notably highlighted strong significant correlations with cardiovascular diseases [Ashar et al., 2017; Zhang et al., 2017b]. The existing methods for estimating copy numbers from DNA microarray data have however been shown as well to provide much less accurate estimates (Pearson correlation coefficient less than 50%) and are, at the time, all array-dependent.

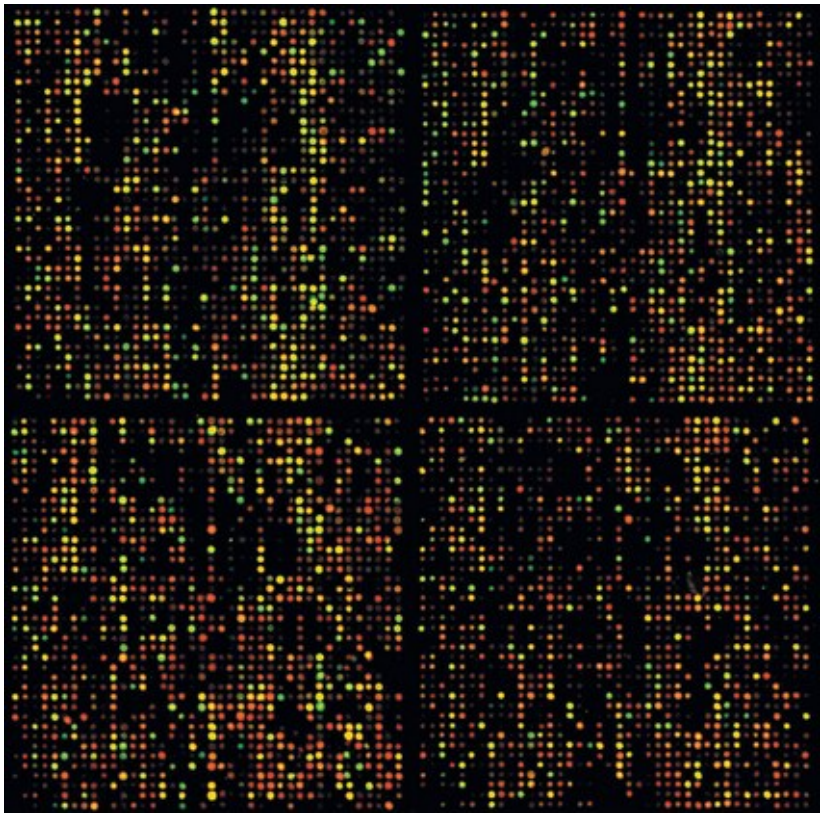

Figure 4: High-throughput DNA microarray plate with fluorescence levels

With the following pipeline, we hypothesise that the measurements of fluorescence intensities across mitochondrial probes can provide accurate information on the amount of mitochondrial DNA present in a sample. Raw mitochondrial probes data is, however, subject to numerous biases and variations across arrays, and can not be used as such. To achieve accurate estimates of mtDNA copy number, we propose a series of pre-processing adjustments whose aim is to standardise data within and across arrays.

3 mtDNA copy number estimation pipeline

In the following section, we present a novel pipeline for estimating mitochondrial DNA copy-number variations between individuals using high-throughput DNA microarray data that can be used over different types of arrays (both from Affymetrix and Illumina) on the sole condition that these arrays contain a sufficient number of mitochondrial probes across the mitochondrial genome.

3.1 Preprocessing

In terms of data, the minimum requirements are per individual raw microarray probes data (including mitochondrial probes); information on each individual's gender and/or age at the time of the study; and, optionally but nonetheless importantly, if the study was led using multiple batches and/or multiple plates, information on each individual's sample batch number and plate number.

The pipeline works in three-steps: preprocessing, clustering, inferring, and, for training purposes and when the data allows it, evaluating. As part of the preprocessing step, single-nucleotide polymorphisms or SNPs (variations of one nucleotide in a DNA sequence) linked to the available mitochondrial probes are selected to avoid potential cross-hybridisation between probes, GC-content in probes is being taken into account and corrected if necessary, as well as the effect of batches and plates on the distributions of signal intensities. The clustering and inferring step consists of clustering all remaining probe signal intensities in a meaningful way, calculating mean probe signal intensities as a counterfactual and estimating mitochondrial DNA copy-numbers via a simple feed-forward neural network. Finally, estimates can be evaluated using standard covariates of mitochondrial DNA copy-numbers, standard measures like the R^2 or the Pearson's correlation coefficient with qPCR or WGS estimates, or using a proportional hazards regression model if data on cardiovascular, neurodegenerative or other phenotypes and diseases is available.

Raw genomic data, and in particular raw probe signal intensities often need a series of adjustments. In the case of high-throughput DNA microarrays, a few of them are cross-hybridisation, GC-content biases and batch and plate effects. Accounting for these problems and trying to play down their effects on the quantitative measures in the available data is a very important step in conceiving accurate estimates of a biomarker because they are likely to distort the estimates and render them less consistent and accurate. To deal with this problem, the first step of our pipeline consists of preprocessing the data to avoid or nullify as much as possible their effects.

3.1.1 SNP selection for cross-hybridisation

As opposed to specific hybridisation (where the two strands of DNA coincide on each base), cross-hybridisation refers to the association between two strands of DNA (in our case, a probe and a DNA molecule from the sample) that are not identical in sequence by complementary base pairing. In copy-number estimations, cross-hybridisation is susceptible to bias the results by badly estimating the fluorescence levels of intensity probes when those are cross-hybridised. Typically, a strand of DNA from the nucleus can pair with a mitochondrial DNA probe imperfectly and create unwanted fluorescence levels for this probe, rendering its measure of the piece of mitochondrial DNA content in the sample noised.

To overcome potential cross-hybridisations, out of all the mitochondrial probes contained on a specific plate, the ones that do not match perfectly the mitochondrial genome or that have a sequence too close to a sequence in the reference human nuclear genome (with frequent SNPs included to allow for differences in the genome) are removed from the analysis.

To do this, all sequences of the mitochondrial DNA probes are BLASTed [Altschul et al., 1990] against the relevant genome reference. The Basic Local Alignment Search Tool or BLAST, which is used here in an improved version, is an algorithm based on a new approach for rapid sequence comparisons that allows finding sequences in a reference DNA (the human one for instance) closest to a studied sequence.

Once sequence similarities are found, all probes that do not match perfectly with the reference mitochondrial genome are removed, so that the only signal intensities kept are directly informative of the amount of mitochondrial DNA in the sample. Then, of the remaining probes, all probes that have off-target above 80% matches with the reference nuclear genome (this threshold can be adjusted) are removed. This allows us to limit the biases introduced by cross-hybridisation and only retain probes that are susceptible to revealing information about the contents in mitochondrial DNA and only in mitochondrial DNA.

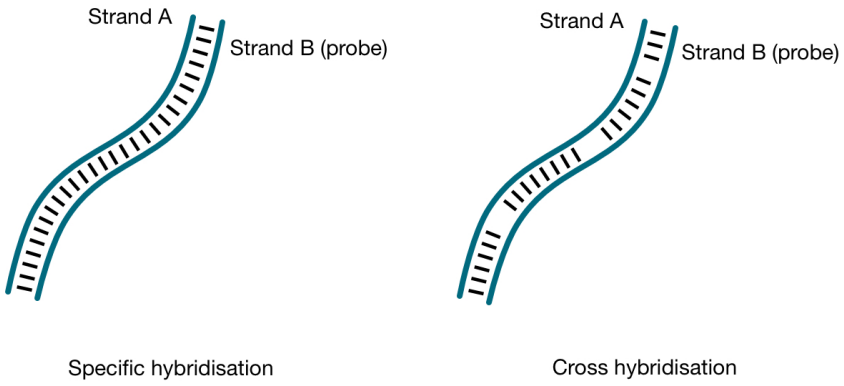

Figure 5: Specific and cross hybridisation

3.1.2 Phasing probe signal intensities

High-throughput DNA microarrays were first introduced to identify small variations in the genome between individuals and determine the effect of these variations on phenotypes and medical conditions. For this reason, the chips typically contain two probes for each small section of the genome, with each probe varying at only one position. The quantitative measures given by the two probes, the intensity read, can then be used to identify which variation the genome of the sample has, and later, its effects on different phenotypes or conditions.

Thus, once all n SNPs have been selected to eliminate potential cross-hybridisation, our data consists of $n \times 2$ probes (two probes for each SNP), with an intensity read associated to each of these probes. All intensity reads should nonetheless not be taken into account, as for each pair of probes, the only meaningful information is carried by the probe whose mitochondrial DNA strand can be combined with the sample's mitochondrial DNA strand. In other words, the data accounts for substitutions of a single nucleotide at a specific position in the mitochondrial genome, and the relevant intensity level is the one of the probe that matches this substitution.

To illustrate this need, in Figure 6, if the mitochondrial DNA strand contained in the sample is the one on the left (SNP with the red T for thymine nucleotide), the only probe that provides us with an informative quantitative measure is the probe A (SNP1), as the intensity from probe B (SNP2) will most likely be linked

to mismatches as its corresponding mitochondrial DNA strand is not contained in the sample's mitochondrial genome.

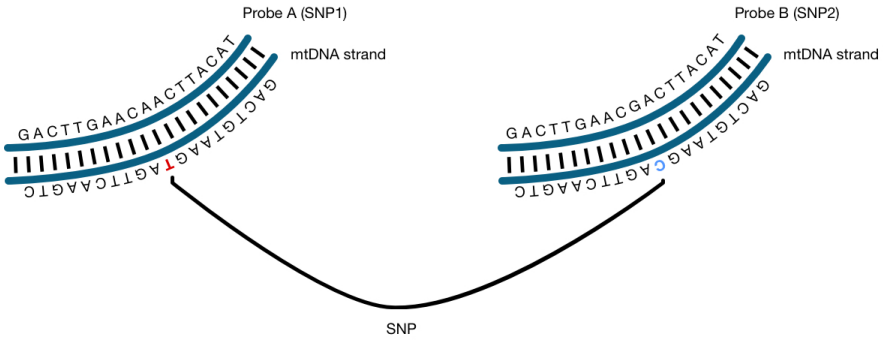

Figure 6: A single-nucleotide polymorphism or SNP

Selecting the correct probe of the two existing probes, which is the same as determining the genotype at each mtDNA position, is called phasing. To find which allele is relevant to our estimations, we use a standard variant calling algorithm for mitochondrial DNA genotyping [Ding et al., 2015].

The algorithm aims to determine, for each given SNP, whether the sample's mitochondrial DNA has one variation or the other. For each SNP, the genotype of the sample can be either A, C, G, T, or multiple alleles can be contained in the same sample (heteroplasmy), in which case the genotype can be A/C, A/G, A/T, C/G, C/T, G/T, A/C/G, A/C/T, A/G/T, C/G/T, or A/C/G/T. The variant calling algorithm determines which of these alleles is in the sample.

Once the correct probe has been identified for each SNP, the remaining n intensity levels compose our main dataset and no further data is removed from the analysis. The data can nonetheless still be biased and further adjustments are needed, in particular GC-content correction and batch and plate effects removal.

3.1.3 GC-content correction

Recently, a number of studies have identified biases linked to GC-content in probes [Benjamini and Speed, 2012; Hildebrand et al., 2010; Romiguier and Roux, 2017]. GC-content is defined as the number of guanine (G) and cytosine (C) nucleotides in a probe. GC-content biases were shown to be positively correlated to the intensity signals of probes, thus rendering studies that focus on measuring fragment abundance within the mitochondrial genome less accurate. In other words, the amount of guanine and cytosine present in a probe is likely to artificially increase the fluorescence intensity levels of the probe, thus creating discrepancies between probes that have different levels of GC-content and making their interpretation less consistent.

In INTERVAL data, for instance, the same can be observed, where the mean phased signal intensities are positively correlated to the GC-content in probes. On Figure 7, each dot represents a single probe, with its relative GC-content (the number of guanine and cytosine nucleotides in the probe divided by the overall number of nucleotides in the probe, that is 50 nucleotides for this type of chip) on the x axis and the mean observed intensity level on the y axis.

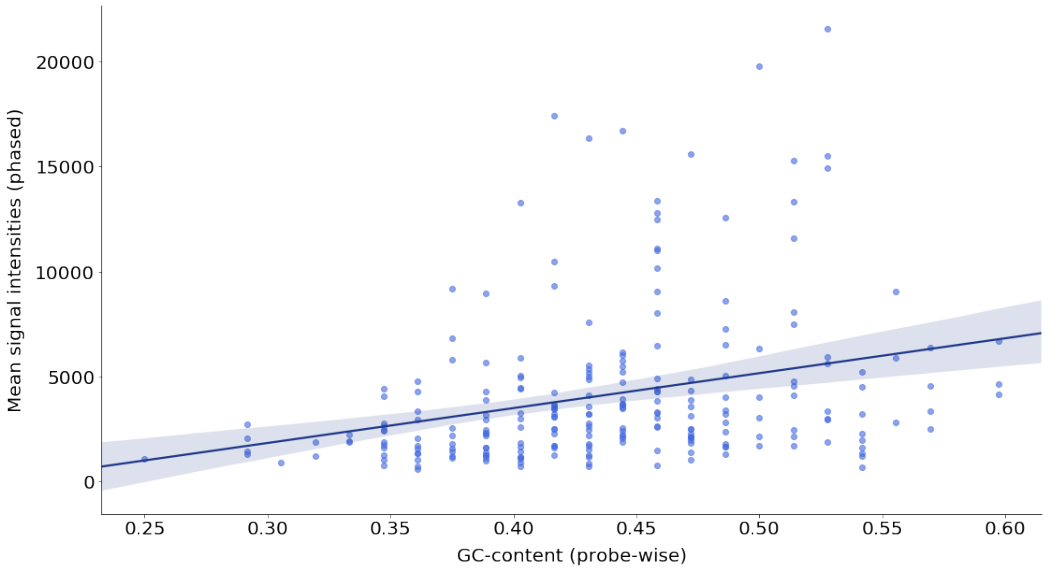

Figure 7: Correlation between GC-content and observed intensity levels

To correct for this, the literature on the subject [Diskin et al., 2008; Benjamini and Speed, 2012] suggets using a simple ordinary least squares model (OLS) model. We correct on a sample-by-sample basis as it was shown to provide better estimates for copy-number analysis. The model used is as such:

$$I_n = \beta_0 + G\beta_1 + \eta \tag{2}$$

with $I_n = (I_{n,1}, I_{n,1}, ..., I_{n,J})$ the vector of signal intensities for all probes for sample n and $G = (g_1, g_2, ..., g_J)$ the vector of relative GC-content in each probe, with g_i the number of guanine and cytosine bases in probe i divided by the overall number of bases in probe i. Using an ordinary least squares model, we have the new (corrected) intensity levels as such:

$$\hat{I}_n^{GC-corrected} = I - (\hat{\beta}_0 + G\hat{\beta}_1) = \hat{\eta} \tag{3}$$

3.1.4 Batch and plate analysis and correction

In large studies like INTERVAL or the data in the UK Biobank, DNA sequencing is often done in multiple batches. Samples are typically analysed in multiple laboratories

and at different times. On top of that, in each batch, samples are usually divided between multiple plates (high-throughput DNA microarrays containing multiple samples). Batches and plates have been widely identified to create shifts in intensity levels distributions due to technical variations [Nygaard et al., 2016] in high-throughput DNA microaarrays, and the data must be used with extreme caution when these effects were not identified.

For instance, in the data from the INTERVAL study, mitochondrial probe signal intensities distributions differed between batches and/or between plates. Keeping them as such can, at best, reduce accuracy in copy-number analyses, and at worst, entirely falsify the results.

Correcting for batch and plate effects can be done in different ways, but a lot of them rely on the randomness of the attribution of batches and plates across samples. In most studies, however, samples were processed and genomes were sequenced using high-throughput DNA microarrays in a way that was first and foremost feasible. For instance, samples were typically processed in the closest capable laboratory from where they were collected. This creates several problems when considering standardising the results from these sequencing procedures.

As a matter of fact, to end up with a uniform and as close to reality as possible distribution of the intensity levels, one has to account not only for the technical variations that create differences between batches and/or plates but also for the differences in the people whose samples compose these batches and plates. Suppose that on average, a batch has higher intensity levels that most other batches in the study. This could very well be due to technical variations in the sequencing procedures used by the laboratories. But it could, however, also be due to differences in the characteristics of the studied population. Differencing these two potentially overlapping effects is what makes this task difficult.

Mitochondrial DNA copy-numbers were indeed shown to be positively correlated to multifarious variables, including sex, age and origin [Zhang et al., 2017b]. If the plate and/or batch is composed of only, or almost only individuals from a specific gender or a specific origin, separating the batch or plate effect between technical variations and population characteristics is much harder and, if this is pushed to the limit, impossible. As a matter of fact, if a batch is composed of samples from female patients only, correcting for potential technical variabilities with other plates without knowing how much the gender has affected the results cannot be done. To try to correct for these differences, we propose a two-step procedure:

(i) analysing batches and plates distributions:

   (a) analyse the distributions of probe signal intensities using Kruskal-Wallis H tests to identify potential differences between batches and plates

   (b) analyse the randomness of individuals distributions across batches and plates using a series of covariates such as gender, age or origin in Pearson's chi-squared tests (i.e. checking whether individuals have been randomly attributed to batches and plates, which is most often not the case)

(ii) correcting for batch/plate effects on probe signal intensities if possible and if needed:

   (a) using across-batch or across-plates usual standardisation if the population was randomly distributed within batches and plates

   (b) using a Generalized Linear Model if (a) fails but if most batches and plates contain sufficient individuals from each gender, age group and/or origin to run the model

   (c) if both (a) and (b) fail, for instance, if for each batch, all samples of the batch come from individuals of a single origin and if this origin changes between batches, then the data of this specific batch is likely not to be exploitable for copy-number estimation

In the INTERVAL study, for instance, the distributions of probe signal intensities were shown to vary between batches and plates using Kruskal-Wallis H tests. Here, we show the results of the test for a randomly selected probe (Affx-92047861). The null hypothesis (H_0) of the Kruskal-Wallis H test is that samples originate from the same distribution, while the alternative hypothesis (H_1) is that samples do not originate from the same distribution. In the case of this probe, the H statistic was of value $H = 6106.61$ (p-value < 0.0001), thus meaning that it is very likely that the samples did not originate from the same distribution, and that there exist significant differences between the batches distributions. Figure 8, which plots the distribution of probe signal intensities for each batch for this probe allows us to visually confirm the differences in distributions across batches.

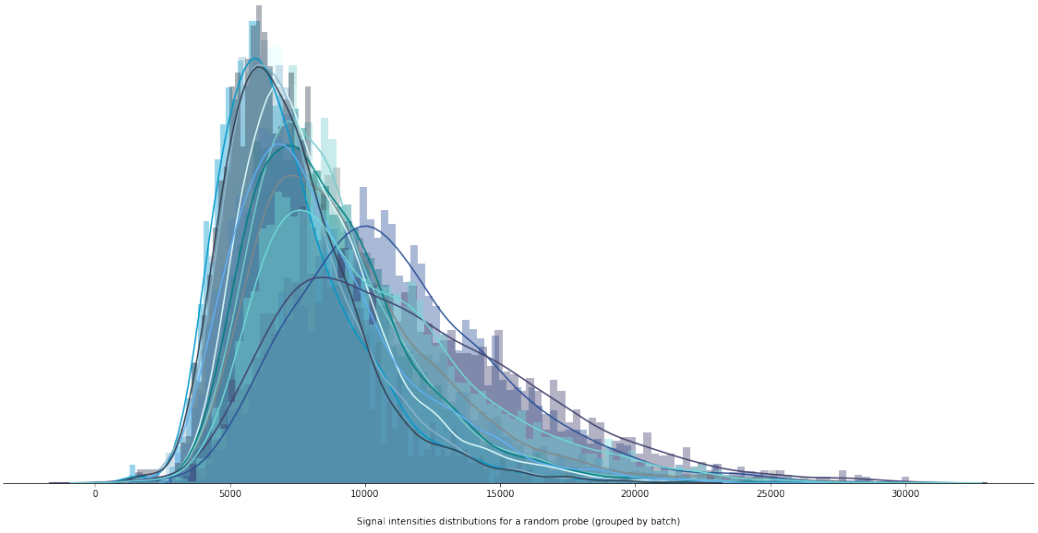

Figure 8: Probe signal intensities distributions across different batches

The difference in distributions raises the potential problem of the non-randomness of individuals distributions across batches and plates. For this, the results regarding the probes of the INTERVAL study suggested that individuals were not randomly allocated to batches and plates of the study. The Pearson's chi-squared tests, whose

null hypothesis (H_0) states that there exists a difference between the distributions and whose alternative hypothesis (H_1) is that the distributions are the same indicated for all probes that the distributions varied from one batch to another. For the probe Affx-92047861, the χ^2 statistic for the analysis of association between batches and gender was $\chi^2 = 23.417$ (p-value $= 0.0053$) and the χ^2 statistic for the analysis of association between batches and age was $\chi^2 = 580.395$ (p-value < 0.00001).

Correction for batch and plate effects can therefore not be made using usual standardisation, as this would decrease or nullify the effect of gender or age on intensity levels. As a result, and after verifying that batches and plates had high-enough proportions for each gender and age group, we chose to correct for batch and plate effects using a GLM, as suggested by the literature [Nygaard et al., 2016].

Before running the Generalised Linear Model, and for it to be more susceptible to work properly, we applied transformations to the probes intensity levels. In the following, we will denote by $I_j = (i_{1,j}, i_{2,j}, ..., i_{N,j})$ the vector of signal intensities of a probe j for all N sample. First, we applied a log-transformation to correct for right-skewness in the distributions:

$$\log(I_j) = (\log(i_{1,j}, \log(i_{2,j}, ..., \log(i_{N,j}) \tag{4}$$

Then, for convenience in future comparisons between arrays and visualisation purposes, and because the key problem does not lie in the truth value of the copy-number but in comparing this value between individuals, we scaled all intensities:

$$I_j^{\text{new}} = (i_{1,j}^{\text{new}}, ..., i_{N,j}^{\text{new}}) \mid \forall k \in [1, n] : i_{k,j}^{\text{new}} = \frac{\log(i_{k,j}) - \min(\log(I_j))}{\max(\log(I_j)) - \min(\log(I_j))} \tag{5}$$

In Figure 9 is plotted the distribution of probe signal intensities for each batch for the randomly selected probe, once transformed for skewness and scaled. The differences in distribution across batches are still visible, but visibly less skewed, which directs us towards using the gaussian family in the GLM.

Batch and plate effect are then corrected using a Generalised Linear Model (GLM) of Gaussian family with log-link function, which allows us to keep gender and age effects in the data whilst correcting for technical variations:

$$\log(I) = X_{BP}\beta_{BP} + X_{AG}\beta AG + \eta \tag{6}$$

with X_{BP} the sparse matrix of batch IDs and plate IDs for all samples and X_{AG} the matrix of genders and ages for all observations.

The resulting level intensities are derived as such:

$$\hat{I}_{\text{batch/plate-corrected}} = \log(I) - X_{BP}\hat{\beta}_{BP} = X_{AG}\hat{\beta}_{AG} + \hat{\eta} \tag{7}$$

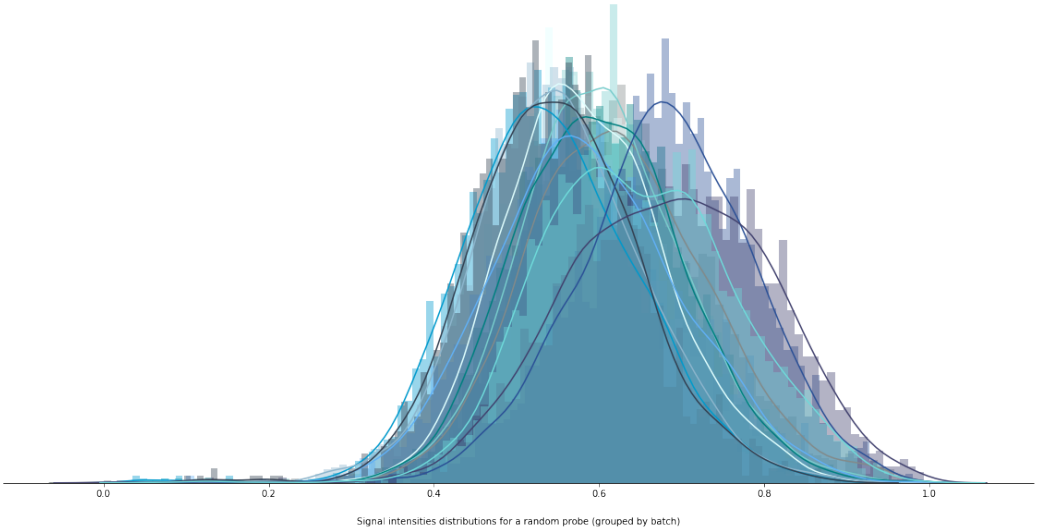

Figure 9: Probe signal log-scaled intensities distributions across different batches

In Figure 10 are plotted the log-scaled probe signal intensities once correction for batch and plate effect has been applied. As can be seen, some variations between batches and plates remain because each batch and plate do not contain as many individuals from the same gender / the same age group, but the distributions are much more consistent with one another than the raw distributions were.

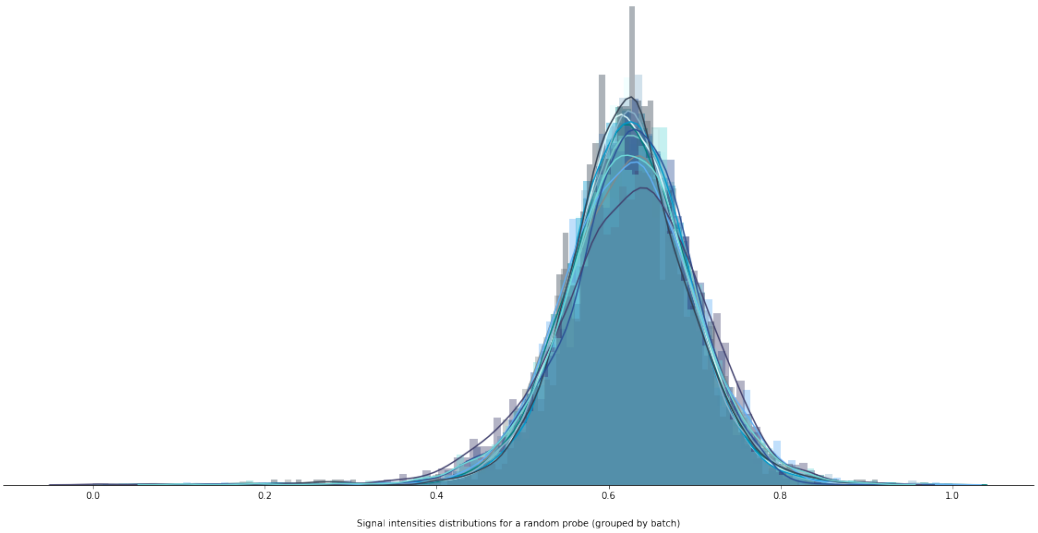

Figure 10: Probe signal corrected intensities distributions across different batches

3.2 Clustering and inference

Once the data is preprocessed and cleaned for further analysis, our pipeline includes a two-step procedure: first, clustering signal intensities in a biologically meaningful way to allow for comparisons and reproducibility on different microarrays with different probe-sets; secondly, an estimation of mitochondrial DNA copy-numbers using either the mean as the predictor or, if either qPCR or WGS estimates are available, a learning algorithm based on a feed-forward neural network.

3.2.1 Probe signal intensities clustering

Recent approaches to estimating mitochondrial DNA copy-numbers using high throughput DNA microarrays have limited themselves to a single study, mostly because of the lack of compatibility between them. The current state-of-the-art in this task [Pankratz, 2018] was for instance designed for the Affymetrix SNP 6.0 chip. Probe-sets differ between most studies because using different SNPs allows researchers to discover new genetic associations with known diseases.

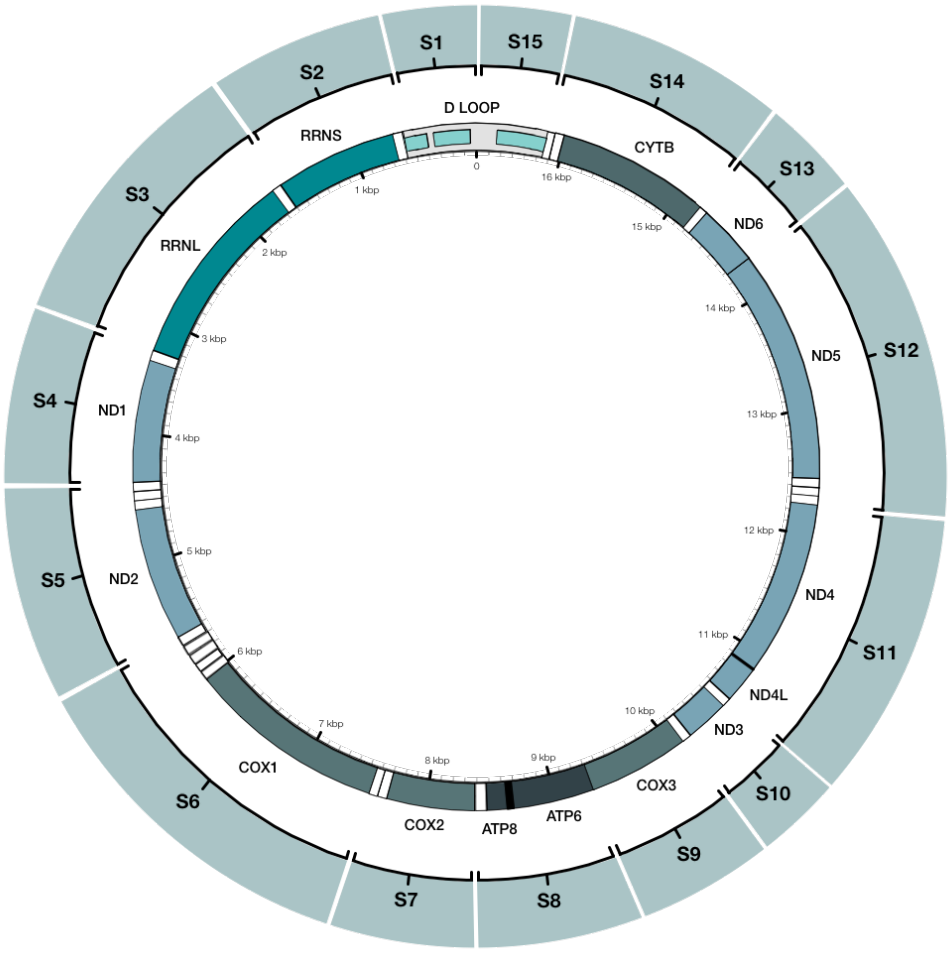

Figure 11: Proposed mtDNA bins for probe intensity levels

The variety of arrays used is therefore particularly valuable when doing genome-wide association studies (GWAS), but this diversity is often limiting for copy-number evaluation. Our approach aims to circumvent this limit by clustering probe signal intensities given the positions of the SNPs on the mitochondrial genome. The aim here is to obtain standardised meaningful features to predict mitochondrial copy-number. To achieve this, we binned the signal intensities in a series of mtDNA clusters derived from the mitochondrial genome reference map and tailored to always contain a few probes.

Figure 11 displays these bins and the associated mitochondrial genome regions. When running the pipeline, the intensity level associated with each bin is the mean of all the intensity levels of all probes whose sequences fall in it. In the (unlikely) case of a probe whose sequence is overlapping two bins, the intensity level associated to the bins are weighted averages of all probes whose sequence fall in it, with the probes whose sequence are only partially included in the bin weighted down proportionally.

The resulting 15 intensity values are common with all high-throughput DNA microarrays as long as the reference mitochondrial genome is not changed. This allows to always have 15 relevant features to estimate mitochondrial DNA copy-numbers and the variations in mitochondrial DNA copy-numbers between individuals. Furthermore, the features are likely to each contain precious information on the amount of mitochondrial DNA in the sample, as it was shown that different parts of the mitochondrial genome have different properties when it comes to measuring the mitochondrial DNA copy-number [Bahreini et al., 2016].

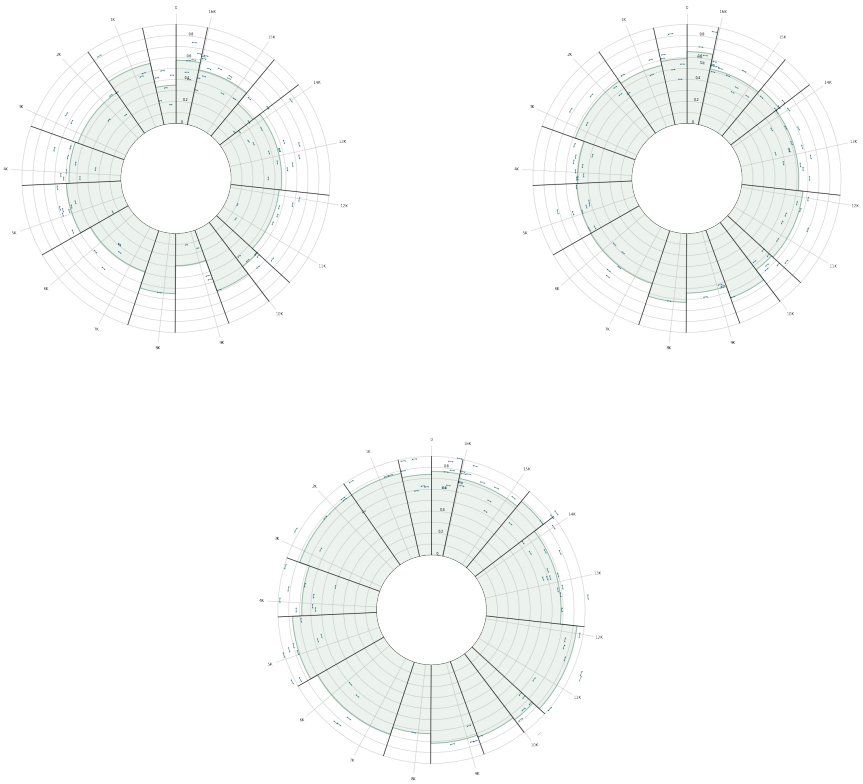

Figure 12: Different intensity levels between samples in the INTERVAL study

Following this procedure, and as is shown in Figure 12, individuals from the INTERVAL study were shown to have different and consistent intensity values for each of the bins. In the Figure, the small blue lines represent the probe sequences along the mitochondrial genome and their distance from the centre their intensity values, and the filled lines the weighted average intensity value for each section:

3.2.2 A simplistic estimate: the mean

The first quick-and-dirty measure that can potentially be used for estimating mitochondrial DNA copy-numbers is a simple weighted average of the signal intensities of each of the bins described in the last subsection, or, in a more straightforward way, the mean of all pre-processed probe signal intensities over the whole mitochondrial genome [Zhang et al., 2017a].

$$\text{mtDNA CN} = \frac{1}{J}\sum_{j=1}^{J} i_{n,j} \tag{8}$$

As a matter of fact, as we showed, mitochondrial DNA copy-number is only a measure of the amount of mitochondrial DNA in a cell. The rationale behind using the mean of all probe signal intensities is that the higher the probe signal intensities are, the more mitochondrial DNA will have interacted with the probe, thus the higher fluorescence levels and higher intensity signals. Averaging over all remaining probes thereby allows to measure, on average, how much mitochondrial DNA interacted with the probes, thus giving an insight into how much mitochondrial DNA was present in the studied individual sample.

3.2.3 Neural network estimates

As was said previously, the literature on the mitochondrial DNA copy-number has identified that different sections of the mitochondrial genome do not play the same role and are not all as likely to allow us to estimate mitochondrial DNA copy-numbers [Zhang et al., 2017a]. The reason behind this is yet to be answered, but could very much be linked to both the different function of the genes on the mitochondrial genome, as well as yet unaccounted for technical variabilities and biases in measuring the probes intensity levels (much like probe-level intensities have for long not been corrected according to their GC-content). For instance, it is known that the D-loop section intensity levels across probes show much more variance than other sections of the mitochondrial genome.

To account for these measures, we tested multiple learning procedures, of which a simple two-layer feed-forward neural network proved to be the most efficient in getting results as close to Whole-Genome sequencing estimates as possible. Deep multilayer feedforward neural networks, which are also called multilayer perceptrons (MLPs) are typically composed of one input layer (the independent variables in an ordinary least squares regression setting), hidden layers and an output layer. In the case of a

regression problem, the output layer typically contains a single node which, once the network has been trained, is the estimate.

The aim of training the neural network is to approximate a function f by defining and recursively estimating weights through backpropagation in order to result in the best function approximation. In the case of a multilayer perceptron, this function is composed of many sub-functions that are chained together. For instance, in our case with two hidden layers, the function f would typically take the form $f(x) = f_3(f_2(f_1(x)))$ with $\forall n \in \{1, 2, 3\}$, each f_n a function that performs an affine transformation of a linear sum of the outputs from the previous function (or the original inputs in the case of f_1). In the case of multilayer perceptrons, the Universal Approximation Theorem (UAT) indicates that any continuous functions on compact subsets of $\mathbb{R}^n$ can be approximated. All these functions traditionally take the form:

$$f_n(x) = g(wx^T + b) \tag{9}$$

with g the activation function (a function that is added into the neural network in order to help it network learn complex non-linear patterns in the data), w the weights, x the the outputs from the previous function (or the original inputs in the case of f_1) and b a bias.

Once defined, the multilayer perceptron is trained in three steps than can be repeated any desired number of times (the epochs):

(i) first, after having initialised the weights (most often to random values), passing the inputs in every layer to obtain the output

(ii) then, calculating a loss function, for instance the quadratic loss, which computes the mean square error between the estimates and the true values, and which is set to evaluate *how wrong* the model is, that is how distant the current output is from the desired output

(iii) and finally, in this last step that is actually training the mode, once the loss is known, backpropagating it and updating the weights of the model accordingly (typically using gradient descent or an optimisation algorithm like Adam or Rectified Adam)

In our case, we used Adam as our optimisation function, the quadratic loss between the mitochondrial DNA copy-number estimates Whole-Genome sequencing and the ones from the neural network as our loss function, and the Rectified Linear Unit function or ReLU as our activation function:

$$\text{ReLU}(x) = \max(0, x) \tag{10}$$

Finally, as is detailed in the third section of this report, hyper-parameters like the learning rate and the batch size were selected using a validation dataset disjoint from the actual test dataset.

3.3 Pipeline summary

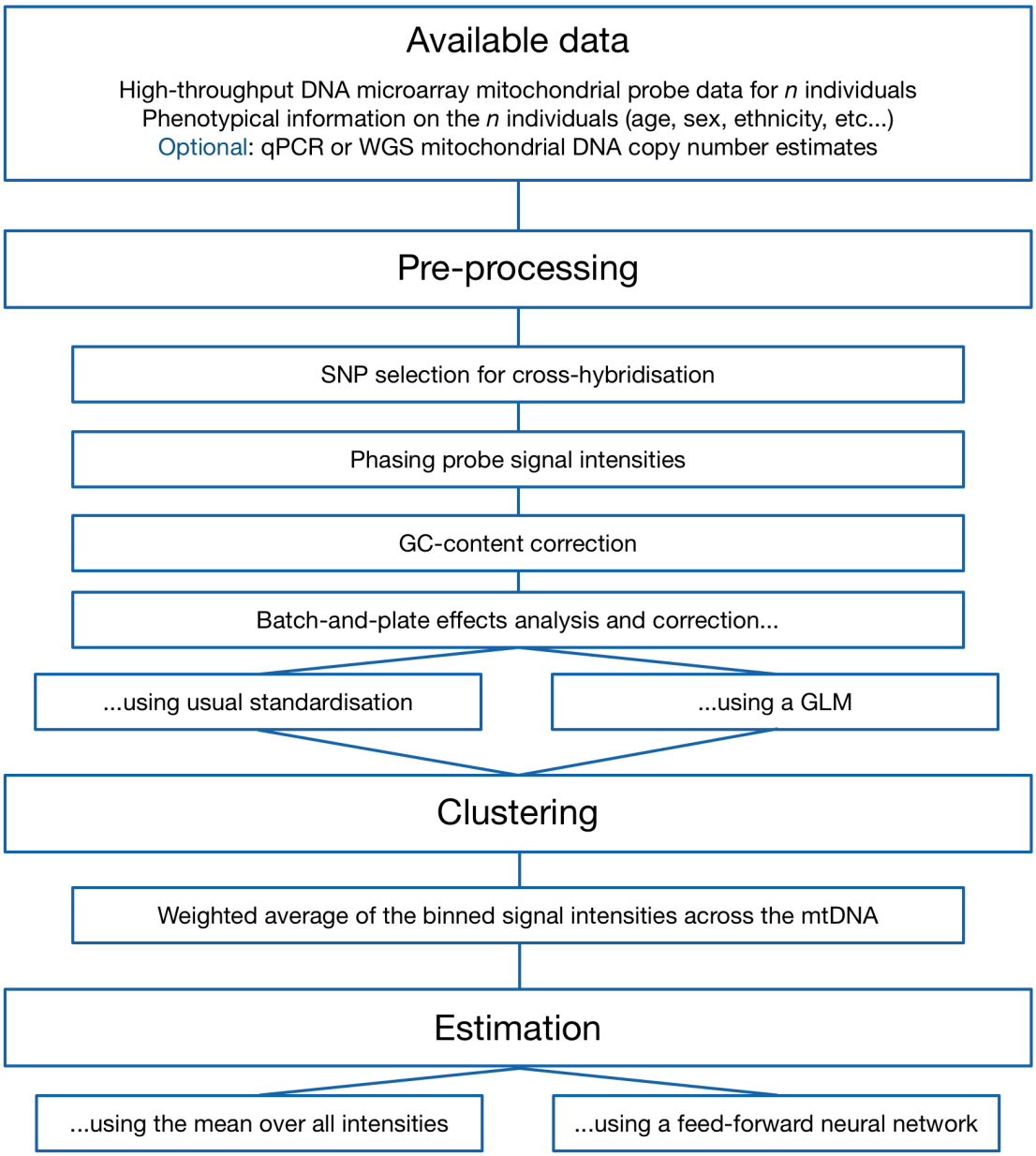

Figure 13: Pipeline summary

3.4 Limits of the pipeline and caution on interpretation

While the pipeline presented here aims to achieve better results in terms of estimation, it is safe to bear in mind that estimating mitochondrial DNA copy-numbers using high-throughput DNA microarrays is, at the moment, still very much a work in progress and an active research topic, and, most importantly, far from being used in a clinical setting.

High-throughput DNA microarrays were not designed to allow copy-number estimations, and the current interest in using them for such tasks and especially for estimating mitochondrial DNA copy-numbers is only the result from their low cost per sample as compared to Whole-Genome sequencing or qPCR, and the renewed interest in mitochondria and their DNA.

While the results presented in the next section show that a positive, highly significant correlation coefficient can be found between mitochondrial DNA copy-number estimates from this pipeline and estimates from Whole-Genome sequencing, it is safe to say that these rely for now on very fragile conditions. As a matter of fact, while studies on genotypes and the links between alleles and medical conditions can very much be achieved despite biases due to GC-content, and batch and plate effects (the only task being to identify the allele in the sample), studies regarding copy-numbers depend crucially on the biases to be corrected in such a way that will make the data uniform across all samples. Nonetheless, as was shown regarding similar arrays that the ones we used, correcting for these biases can prove to be a very difficult task, if not in some cases an unachievable one [Nygaard et al., 2016].

Besides, learning algorithms such as a deep feed-forward neural network depend a lot on the amount of data available. However, at the time, studies with multiple thousands of samples that have been sequenced both using high-throughput DNA microarrays and Whole-Genome sequencing are scarce, and studies that have mitochondrial DNA copy-number estimations using qPCR for thousands of samples simply do not exist, making the reproducibility of such pipeline particularly weak. Estimates gathered from algorithms like this one should therefore be taken with caution and, perhaps, only be interpreted as relatively good proxies of the novel biomarker that is the mitochondrial DNA copy-number rather than accurate estimates of it.

4 Discussion on the results

Evaluating the quality of mitochondrial DNA copy-number estimates relies heavily on knowing the true value of the copy-number for a large proportion of samples in the data. Unfortunately, and pretty much for the same reason as the one that originated this research (cost), qPCR estimates were not available. As a matter of fact, at the time of the estimations, there was no known study with more than 50 qPCR mitochondrial DNA copy-number estimates and high-throughput DNA microarray data.

Subsequently, we thought of three ways to get insights into how good our estimates were despite not knowing the true value that we aim for. First, as we have highlighted in the first section, Whole-Genome sequencing estimates of the mitochondrial DNA copy-number were shown to be particularly consistent and accurate.

While qPCR, which we would define as the gold-standard, is not available, a fair way to check the validity of our results is to compare WGS estimates which, as was shown, are close to qPCR estimates, to our new high-throughput DNA microarray estimates. This is possible for us through the 1000 Genomes study, which consists of more than two thousand samples with both Whole-Genome sequencing and

high-throughput microarray available. Unfortunately, the 1000 Genomes study did not contain information on batches and thus batch or plate effects were not corrected for in the following section.

Then, for the INTERVAL study, as Whole-Genome sequencing data was not available, comparing estimates from two different sources was not possible. Results using the data from the INTERVAL study should thus be interpreted with lots of caution. Nonetheless, we also know that mitochondrial DNA copy-number (calculated from qPCR estimates) is highly correlated to numerous known variables, including age, sex and numerous blood cell counts. To try to assess whether our estimates aimed in the right direction, we computed the Pearson correlation coefficients between them and these known correlates and compared them to the literature on the subject. Finally, as we underlined before, mitochondrial DNA copy-number estimates are also highly correlated to several medical conditions, including cardiovascular and neurodegenerative diseases. One final way to assess the quality of our estimates would thus lie in using them directly in a proportional hazards model. Unfortunately, data from the UK Biobank was not made available to us in time for this to be done.

4.1 Comparison with Whole-Genome Sequencing estimates

One way to measure the quality of high-throughput DNA microarray estimates without knowing the value of qPCR estimates is to compare the results we obtained with results obtained from Whole-Genome sequencing estimates.

To do this, we estimated the mitochondrial DNA copy-number using WGS data from the 1000 Genomes study using fastMitoCalc [Qian et al., 2017]. fastMitoCalc is a program that uses all of the mitochondrial genome reads and a small subset (less than 1%) of the autosomal genome reads in the sample to estimate both mitochondrial and autosomal DNA coverages accurately.

The metric to assess the accuracy of our results depends a lot on the metrics used by the literature on the same subject. While in many cases where regression tasks are done using machine-or-deep learning algorithms, the mean squared error (MSE), root mean squared error (RMSE) and mean absolute error (MAE) are useful when comparing algorithms together, these cannot be applied in this case and have not been used by the literature. As a matter of fact, there is no standard dataset for this specific task and researchers typically test different data given their availability. Given the fact that different microarray chips usually yield intensities in different ranges and with very different distributions, lower MSE/RMSE/MAE could be achieved thanks to the distribution of the data rather than by real improvements in the algorithms. As of today, the metric commonly used by the research articles on the subject is the Pearson (or sometimes Spearman) correlation coefficient, and the R-squared [Longchamps et al., 2020]. At the time of writing, MitoPipeline, a subprogram of Genvisis for generating mitochondrial copy-number estimates from DNA microarray data is giving the best results, with a Pearson correlation coefficient of 50% [Pankratz, 2018].

In the following part, we use a standard OLS regression to visualise the results

and comment them with both the R-squared resulting from this regression and the Pearson correlation coefficient between the WGS estimates and the high-throughput DNA microarray estimates. We show that while the mean estimates do not perform well, the neural network estimates on the test dataset perform better than the current state-of-the-art for the task.

4.1.1 Mean estimates

The rationale behind using the mean as a predictor of the mitochondrial DNA copy number is that probe intensities reflect, on average, the amount of mitochondrial DNA in the sample [Zhang et al., 2017a].

Figure 14 shows indeed that there is a positive and significative relationship between the mitochondrial DNA copy-number mean estimates from high-throughput DNA microarrays and the estimates derived from Whole-Genome sequencing data. In our case, with $n = 2150$, The Pearson correlation coefficient between the two is 33.58% and the R-squared associated to the OLS regression of the high-throughput microarray estimates on the WGS estimates is $R^2 = 0.11$.

While these results are insufficient for any further analysis and cannot be used as such in a research study on the relationship between mitochondrial DNA copy-number and cardiovascular or neurodegenerative diseases, they shed light on the fact that the quantity estimates from high-throughput DNA microarrays reflect to a certain extent and with potential biases the same quantity measured by WGS estimates, that is the amount of mitochondrial DNA in a sample/cell.

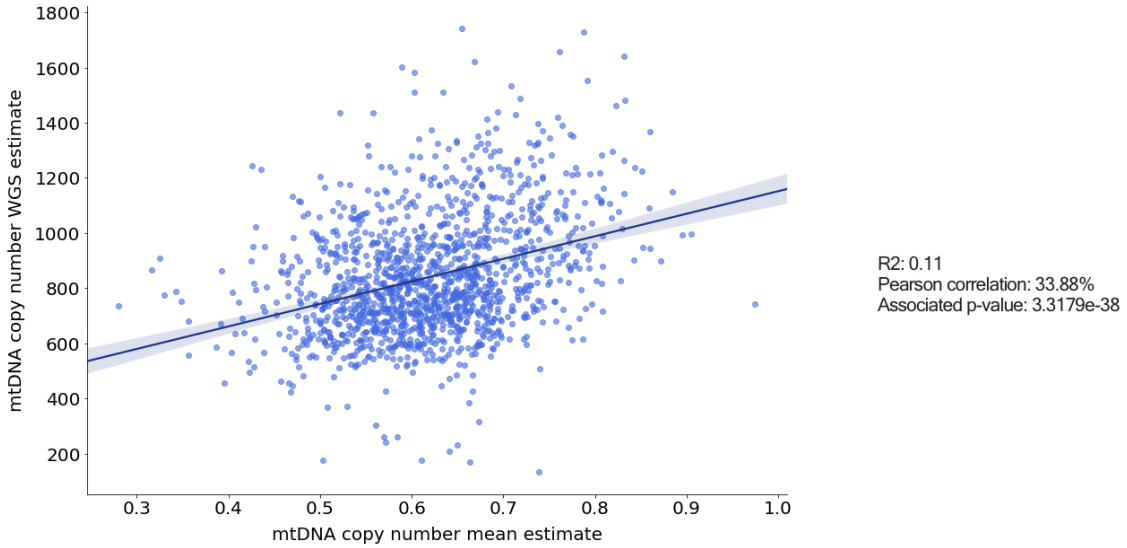

Figure 14: OLS regression of WGS estimates on mean estimates

There are, however, two major drawbacks of estimating the mitochondrial DNA copy-number using the mean. The first, as was shown in the second section, is that this does not take into account the potential variations in predictive power between different sections of the mitochondrial genome.

In particular, it was shown that sections of the genome do not provide information with the same accuracy as others and that these should likely not be all used in the same way when estimating mitochondrial DNA contents in a sample [Zhang et al., 2017a].

A secondary drawback is that the mitochondrial DNA contents are estimated without taking into account the overall DNA contents of the sample. For technical reasons, some samples are indeed likely to return, on average and over the entire nuclear and mitochondrial genomes, higher intensities than others. Intensities from mitochondrial DNA probes may thereby be higher not because the amount of mitochondrial DNA in the sample is higher, but because the sample's interaction with the high-throughput microarray has, on average, worked better than for another sample.

4.1.2 Neural network estimates

As a result of these limitations and in order to obtain more accurate estimates of the mitochondrial DNA copy-number, we also trained our feed-forward neural network using the Whole-Genome sequencing estimates as the dependant variable.

We used two types of independent variables or inputs in the network. First, the mean section intensities obtained after binning the probes intensity levels (see Figure 11), which consist in 15 variables that we hypothesised give different types of information on the level of mitochondrial DNA content in the sample. Then, To take into account potential biases due to differences in the overall DNA contents of the sample, we performed a PCA on all autosomal probe intensities and kept the first three principal components as inputs. Only the three first principal components were kept as these have been shown to carry most of the information regarding the amount of autosomal DNA in the sample. In the end, the inputs and desired output of the neural network are the ones in Figure 15.

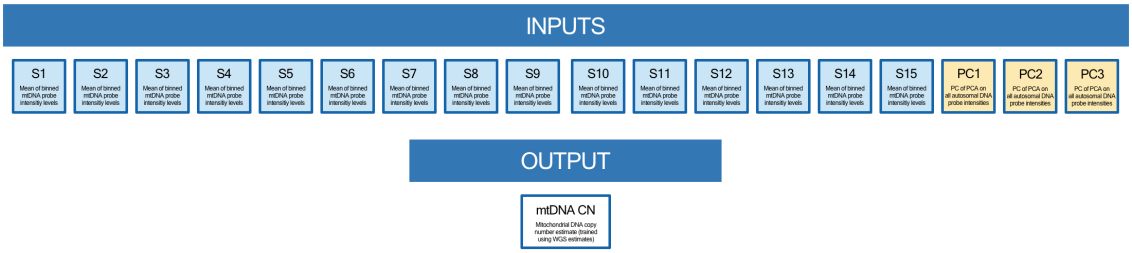

Figure 15: Inputs and dependent variable of the neural network

To train the neural network, we split the data randomly in three sets:

(i) *train*: a dataset of 1400 samples that are used to train the network, i.e. the data that is used to adjust the weights by back-propagation

(ii) *validation*: a dataset of 350 samples that is used to evaluate the model while adjusting the hyper-parameters and correcting for potential under-fitting or over-fitting problems, i.e. the data used to provide information on how accurately

the regression task is achieved and that is disjoint from the actual test dataset, thus making sure that the hyper-parameters are not being optimised for the test set

(iii) *test*: a dataset of 350 samples used to evaluate the final model and provide us with unbiased results regarding the final model

Hyper-parameters like the number of hidden units or the learning rate were therefore chosen using the validation dataset. On top of that, given that our training dataset was small (we had not been given access to the UK Biobank datasets which would have contained much more data), and to avoid overfitting, we added Gaussian noise to the inputs at each batch and early-stopped the model after 300 epochs.

Once trained, the results of the feed-forward neural network can be seen in Figure 16, which shows OLS regressions between the neural network mitochondrial DNA copy-number estimates and the ones from Whole-Genome Sequencing data for each of the train, validation and test datasets.

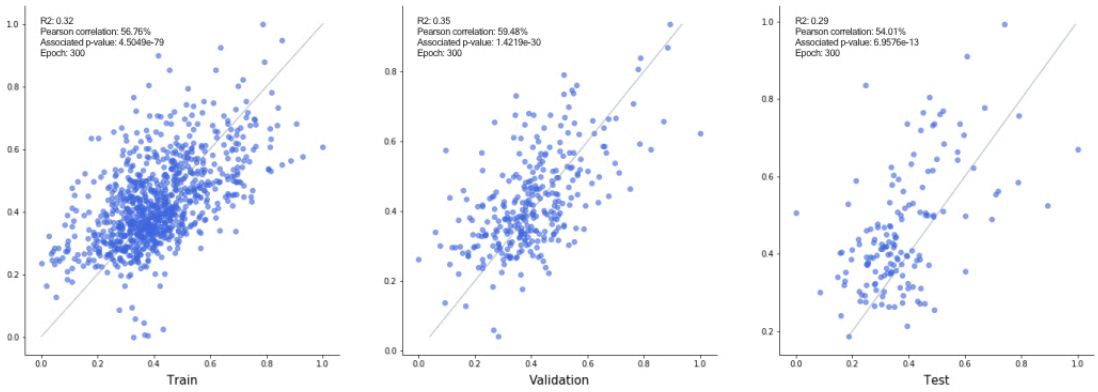

Figure 16: OLS regression of WGS estimates on neural network estimates

As is shown, the test dataset estimates perform better than the estimates from the mean of all probe intensities, with $R^2 = 0.29$ and a Pearson correlation coefficient equal to the squared root of the R^2, that is 54.01%. While this amount of data, both for the training dataset and the testing one, is insufficient to prove the current accuracy of this method, these results do slightly better than the current state-of-the-art algorithm (MitoPipeline from Genvisis, which achieves a 50% Pearson correlation coefficient).

Finally, as the training data is obtained from a pipeline which can theoretically be applied independently from the probes present in the high-throughput DNA microarray, and which results in similarly distributed input variables, this method aims to be reproducible for larger datasets and, given that there isn't too much variation between the probe intensity distributions, the training weights can be improved using datasets from other studies. Unfortunately, at the time of the study, no other datasets containing both high-throughput DNA microarrays and Whole-Genome sequencing estimates were available, and we did not test this potential for across-study or across-arrays compatibility.

4.2 Correlations with standard covariates

As is said earlier, the true value of mitochondrial DNA copy-number was not available to us and could only be approached by the proxy of Whole-Genome sequencing estimates. As a supplementary way to get first sight of the quality of high-throughput DNA microarray estimates, we assessed them with regards to standard mitochondrial DNA copy-number covariates.

Mitochondrial DNA copy-number has indeed been shown by the literature to be associated with numerous standard phenotypical and biological variables. INTERVAL data allowed us to test both the usual association with sex and with standard blood cell counts. In similar studies, the mitochondrial DNA copy-number was shown to be higher in females than males [Longchamps et al., 2019; Hulgan et al., 2019], and to be associated with white blood cells count and with platelet count [Longchamps et al., 2019; Hulgan et al., 2019; Hurtado-Roca et al., 2016; Shen et al., 2008]. Here, we computed the Pearson correlation coefficients between mitochondrial DNA copy-number mean estimates and covariates. While it is important to bear in mind that these results can by themselves in no way confirm that the value we are estimating is representative of the mitochondrial DNA copy-number, they offer clues in believing this is the case and, potentially, in leading this study further and with additional data.

Variable	Pearson correlation (associated p-value)
Sex	-3.28% ($p = 1.06 \times e^{-11}$)
White blood cells count	10.28% ($p = 1.84 \times e^{-95}$)
Red blood cells count	2.23% ($p = 7.54 \times e^{-06}$)
Platelets count	-3.21% ($p = 2.45 \times e^{-10}$)
Neutrophils percentage	8.79% ($p = 5.33 \times e^{-68}$)
Eosinophils percentage	-1.89% ($p = 0.00015$)
Basophils percentage	-3.15% ($p = 2.97 \times e^{-10}$)
Granulocytes count	11.44% ($p = 1.69 \times e^{-113}$)
Monocytes count	4.69% ($p = 1.71 \times e^{-20}$)
Lymphocytes count	1.27% ($p = 0.01187$)

In the case of the INTERVAL study, the mean estimates from high-throughput DNA microarray data show associations with the same covariates, as well as other, weaker associations with blood cell counts (for neutrophils and granulocytes in particular).

While women have repeatedly been shown to have higher mitochondrial DNA copy-numbers than males, the sign of the blood cell count correlations vary from study to study. Nonetheless, it is only fair to note that our estimates show a positive

correlation coefficient while most studies with estimates from high-throughput DNA microarray data show negative correlation coefficients [Longchamps et al., 2019]. This difference in sign that does not corroborate most studies that have found a similar association was investigated but not explained at the time of writing this report.

4.3 Using the mtDNA copy-number estimates in practice

The main use of mitochondrial DNA copy-number estimates is, at the moment, one of research about its potential use as a biomarker for a variety of medical conditions.

This is typically done in studies that share both genomic data and in this case high-throughput DNA microarray data, as well as variables regarding both the individuals' age, sex, ethnicity, phenotype, past medical conditions and life habits (smoking, drinking, etc.).

The UK Biobank datasets contain a wide variety of variables including the medical events in the life of patients and the dates at which they happened, thus allowing for such study. In this case, survival analysis is possible and possibly gives valuable insights into the relationship between a biomarker and the studied medical conditions. These studies often use Cox's proportional hazards models [Zhang et al., 2017b; Pyle et al., 2015].

Cox's proportional hazards models are semi-parametric models for studying durations and survival in medical studies. They work by evaluating the time elapsed before an event occurs. Historically, in the formulation proposed by the British statistician David Cox, this event is the death of the individual being studied. In our case, it can as well be the death of the patient, but other events can be taken into such analysis as well, as cardiac arrest, age of onset of neurodegenerative diseases, kidney failure, etc.

In the following, we are interested in T, a positive random variable modelling the amount of time before the studied event, i.e. the number of days separating the origin time of the study and the time when the individual experiences a medical condition or death. Here, we denote by F the cumulative distribution function (CDF) function of T, meaning that we have $\forall t > 0$:

$$\mathbb{P}(T < t) = F(t)$$

Using these notations, we can write the survival function as such, i.e. the probability that the patient *survived* or did not experience any medical event of the type of those studied at time t:

$$\forall t > 0, \quad S(t) = \mathbb{P}(T > t) = 1 - F(t)$$

And, likewise, the hazard function, which measures the instantaneous probability of the even occurring at time t, knowing that nothing had happened before t:

$$\forall t > 0, \quad h(t) = \lim_{\eta \to 0} \frac{\mathbb{P}(T \in]t, t + \eta[| T > t)}{\eta} = \frac{f(t)}{S(t)}$$

Cox's proportional hazards model then assumes that the hazard function is composed of both:

(i) a time-dependent basic hazard function $\lambda_0(t)$, also called baseline hazard, corresponding to the instantaneous risk of an event occurring when all variables of interest are equal to zero. This basic hazard is never specified, and can therefore take any form. It is worth noting and remembering that no attempt is made to estimate it in the following, thus making the Cox proportional hazard model semi-parametric

(ii) a secondary hazard function $e^{\beta_1 x_{i,1}+...+\beta_k x_{i,k}}$ defined for any individual who depends only on the studied covariates (and therefore not on time)

One can then write, and develop this way:

$$\begin{aligned} \forall t > 0, \forall i, \quad & h_i(t) = \lambda_0(t) e^{\beta_1 x_{i,1}+...+\beta_k x_{i,k}} \\ \Leftrightarrow \quad & \ln h_i(t) = \ln(\lambda_0(t)) + \beta_1 x_{i,1} + ... + \beta_k x_{i,k} \end{aligned}$$

This semi-parametric formulation allows us to look at the instantaneous risk ratio between two modalities of a variable. As a matter of fact, it is worth noting again that we are not trying to estimate the function $\lambda_0(t)$ (which gives the baseline hazard of an event occurring at time t), but rather the ratio of the instantaneous risks of an event occurring for two individuals exposed to different risk factors.

Formally, this yields, by denoting $\mathrm{R}_{i,j}$ the ratio of the instantaneous risks between an individual i and an individual j which differ only by one modality of one variable:

$$\begin{aligned} \forall t > 0, \forall i, j, \quad \mathrm{R}_{i,j} &= \frac{h_i(t)}{h_i(t)} \\ &= \frac{\lambda_0(t) e^{\beta_1 x_{i,1}+...+\beta_m x_{i,m}+...+\beta_k x_{i,k}}}{\lambda_0(t) e^{\beta_1 x_{j,1}+...+\beta_m x_{j,m}+...+\beta_k x_{j,k}}} \\ &= \frac{e^{\beta_1 x_{i,1}+...+\beta_m x_{i,m}+...+\beta_k x_{i,k}}}{e^{\beta_1 x_{i,1}+...+\beta_{m-1} x_{i,m-1}+\beta_m x_{j,m}+\beta_{m+1} x_{i,m+1}+...+\beta_k x_{i,k}}} \\ &= e^{\beta_m (x_{i,m} - x_{j,m})} \end{aligned}$$

For instance, if the variable X_m is binary, and that it is equal to 1 for the individual i and 0 for the individual j, we have :

$$\begin{aligned} \forall t > 0, \forall i, j, \quad & \mathrm{R}_{i,j} = e^{\beta_m} \\ \Leftrightarrow \quad & \ln(\mathrm{R}_{i,j}) = \beta_m \end{aligned}$$

The instantaneous risk ratio is thus independent with time. In other words, for any time t, the individual i has an instantaneous risk of experiencing the event equal to $\mathrm{R}_{i,j} = e^{\beta_m}$ times that of individual j. This ratio thus allows us to measure the effect of a variable on the overall chances of survival, while also allowing for the fact all individuals are exposed to the risk of such event (death, cardiac arrest, etc.) independently from the variables studied.

Estimating $\beta_1, ..., \beta_k$ is typically done by maximum likelihood, and the coefficients can then be interpreted as follows:

(i) for a binary or dichotomous variable X_m taking two values: 0 if the individual is of type A and 1 if the individual is of type B, we will say that individuals of type A have an instantaneous risk of death/experiencing the studied medical event equal to e^{β_m} times that of individuals of type B

(ii) for a continuous variable X_n we will say that the individuals for whom this variable is s have an instantaneous risk of death/experiencing the medical event equal to $e^{\beta_n s}$ times that of the individuals for whom this variable is 0

Last but not least, it should also be noted that Cox proportional hazards models take into account the fact that the data are truncated or censored, which is the case of almost all medical studies. As a matter of fact, when studying, for instance, the risk of cardiac arrest, not all patients in the study will have experienced it at the time of publication of the dataset nor possibly in their life. While this leaves uncertainty regarding the outcome of certain patients, it is important that the model uses the fact that at the time when the study was stopped (right-censored time), these same patients have not experienced the studied event.

Unfortunately, despite great efforts from our team to speed up the process, some UK Biobank datasets were not made available to us before the end of this internship, and the results of a Cox proportional hazards model on our estimates are thus let for future work on the project.

5 Conclusion

Estimating mitochondrial DNA contents from high-throughput DNA data is, therefore, a difficult task and still an active research question. While high-throughput DNA microarrays have become cheaper and using them in large studies possible, they were not conceived for estimating copy-numbers, and thus fail in allowing researchers to easily measure the amount of DNA contained in samples. The difficulties typically encountered when trying to do so inherently belong to the design of high-throughput DNA microarrays. As a matter of fact, inferring copy-numbers from probe intensity data results implies relying on the idea that intensity fluorescence levels in probes reflect accurately the amount of DNA that interacts with the probe. However, probe intensity levels are subject to multifarious effects that can make their value deviate from that of the amount of DNA they hybridise with.

While numerous biases like the potential cross-hybridisation of probes with the nuclear genome and the effect of GC-content on fluorescence levels have been identified, correcting for them is no easy task. On top of that, high-throughput DNA microarrays are processed in batches and plates and are known to be prone to have their measurements vary due to technical variabilities. While this is not a problem for genotyping tasks, batch and plate effects can highly affect copy-number estimations and can not always be corrected for. Finally, new studies typically use different high-throughput DNA microarrays which allow researchers to look into the characteristics of newly studies genes, but which also create a lot of variabilities that render comparing copy-number estimations across studies difficult.

The pipeline presented in this report aims to cover as much as possible of these potential biases and to correct them. Nonetheless, and despite getting slightly better results than the widely used MitoPipeline software for mitochondrial DNA copy-number estimation, the estimates are far from being good enough to be used in a clinical research setting. It is nonetheless worth noting that the results presented here come from the 1000 Genomes study which contains a wide variety of individuals from different ethnicities and ages, and no information was given on batch and plates, whose effects were therefore not corrected for. Knowing how batches and plates can affect the distribution of nuclear intensities, it is safe to say that the results presented here let room for improvement.

A Glossary

Below is presented a non-exhaustive glossary of biological terms:

Adenosine triphosphate (ATP): Energy-carrying molecule found in the cells of all living things.

Apoptosis: Programmed cell death; deletion of individual cells by fragmentation into membrane-bound particles.

Autosomal: Of any chromosome other than the sex chromosome pair.

Copy-number: Number of copies of the studied genetic molecule in a cell.

Cross-hybridisation: Molecular biology technique that measures the degree of genetic similarity between pools of DNA sequences.

Deoxyribonucleic acid (DNA): Organic chemical of complex molecular structure carrying genetic instructions for the development, functioning, growth and reproduction of all known organisms.

Eukaryote: Cell containing a membrane-bound nucleus with chromosomes of DNA and proteins with DNA transcription inside the nucleus and protein synthesis in the cytoplasm, in contrast to prokaryotes.

Fluorophore: Any of various chemical groups or structural domains that are responsible for the fluorescent properties of a substance.

Genotype: The genetic makeup, as distinguished from the physical appearance, of an organism or a group of organisms.

Heteroplasmy: Presence of more than one type of organellar genome (mitochondrial DNA or plastid DNA) within a cell or individual.

Homoplasmy: Presence of a single type of organellar genome (mitochondrial DNA or plastid DNA) within a cell or individual.

Hybridisation (as in nucleic acid hybridisation): Process of joining two complementary strands of nucleic acids - RNA, DNA or oligonucleotides.

Microarray (as in DNA microarray): Collection of microscopic DNA spots (probes) attached to a solid surface on a solid substrate (usually a glass slide or silicon thin-film cell) that assays large amounts of biological material using high-throughput screening.

Microarray plate (as in DNA microarray): Plate containing multiple wells where each well contains a sample analysed using a microarray.

Mitochondrion: Double membrane bound organelle found in the cytoplasm of eukaryotic cells.

Nucleotide: Organic molecules that serve as the monomer units for forming the nucleic acid polymers DNA (deoxyribonucleic acid) and RNA (ribonucleic acid).

Organelle: Specialised subunit within a cell that has a specific function.

Phenotype: Composite of an organism's observable characteristics or traits, such as its morphology, development, biochemical or physiological properties, behaviour, etc.

Prokaryote: Organisms whose cells have no nucleus or membrane-bound organelle, e.g. bacteria.

Single Nucleotide Polymorphism (SNP): Occurrence of more than one form, as several alleles of a particular gene or winged and wingless forms of the same species.

List of Figures

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410.

Andersson, G., Karlberg, O., Canbäck, B., and Kurland, C. G. (2003). On the origin of mitochondria: a genomics perspective. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 358(1429):165–179.

Ashar, F. N., Zhang, Y., Longchamps, R. J., Lane, J., Moes, A., Grove, M. L., Mychaleckyj, J. C., Taylor, K. D., Coresh, J., Rotter, J. I., et al. (2017). Association of mitochondrial dna copy number with cardiovascular disease. *JAMA cardiology*, 2(11):1247–1255.

Bahreini, F., Houshmand, M., Modaresi, M. H., Tonekaboni, H., Nafissi, S., Nazari, F., and Akrami, S. M. (2016). Mitochondrial copy number and d-loop variants in pompe patients. *Cell Journal (Yakhteh)*, 18(3):405.

Benjamini, Y. and Speed, T. P. (2012). Summarizing and correcting the gc content bias in high-throughput sequencing. *Nucleic acids research*, 40(10):e72–e72.

Cai, N., Li, Y., Chang, S., Liang, J., Lin, C., Zhang, X., Liang, L., Hu, J., Chan, W., Kendler, K. S., et al. (2015). Genetic control over mtdna and its relationship to major depressive disorder. *Current Biology*, 25(24):3170–3177.

Chan, N. C., Salazar, A. M., Pham, A. H., Sweredoski, M. J., Kolawa, N. J., Graham, R. L., Hess, S., and Chan, D. C. (2011). Broad activation of the ubiquitin–proteasome system by parkin is critical for mitophagy. *Human molecular genetics*, 20(9):1726–1737.

Cho, D.-H., Nakamura, T., and Lipton, S. A. (2010). Mitochondrial dynamics in cell death and neurodegeneration. *Cellular and Molecular Life Sciences*, 67(20):3435–3447.

Chu, H.-T., Hsiao, W. W., Tsao, T. T., Chang, C.-M., Liu, Y.-W., Fan, C.-C., Lin, H., Chang, H.-H., Yeh, T.-J., Chen, J.-C., et al. (2012). Quantitative assessment of mitochondrial dna copies from whole genome sequencing. In *BMC genomics*, volume 13, page S5. BioMed Central.

Delbarba, A., Abate, G., Prandelli, C., Marziano, M., Buizza, L., Arce Varas, N., Novelli, A., Cuetos, F., Martínez, C., Lanni, C., et al. (2016). Mitochondrial alterations in peripheral mononuclear blood cells from alzheimers disease and mild cognitive impairment patients. *Oxidative Medicine and Cellular Longevity*, 2016.

Ding, J., Sidore, C., Butler, T. J., Wing, M. K., Qian, Y., Meirelles, O., Busonero, F., Tsoi, L. C., Maschio, A., Angius, A., et al. (2015). Assessing mitochondrial dna variation and copy number in lymphocytes of˜ 2,000 sardinians using tailored sequencing analysis tools. *PLoS genetics*, 11(7):e1005306.

Diskin, S. J., Li, M., Hou, C., Yang, S., Glessner, J., Hakonarson, H., Bucan, M., Maris, J. M., and Wang, K. (2008). Adjustment of genomic waves in signal intensities from whole-genome snp genotyping platforms. *Nucleic acids research*, 36(19):e126–e126.

Fazzini, F., Lamina, C., Fendt, L., Schultheiss, U. T., Kotsis, F., Hicks, A. A., Meiselbach, H., Weissensteiner, H., Forer, L., Krane, V., et al. (2019). Mitochondrial dna copy number is associated with mortality and infections in a large cohort of patients with chronic kidney disease. *Kidney international*, 96(2):480–488.

Fazzini, F., Schöpf, B., Blatzer, M., Coassin, S., Hicks, A. A., Kronenberg, F., and Fendt, L. (2018). Plasmid-normalized quantification of relative mitochondrial dna copy number. *Scientific reports*, 8(1):15347.

Götz, A., Tyynismaa, H., Euro, L., Ellonen, P., Hyötyläinen, T., Ojala, T., Hämäläinen, R. H., Tommiska, J., Raivio, T., Oresic, M., et al. (2011). Exome sequencing identifies mitochondrial alanyl-trna synthetase mutations in infantile mitochondrial cardiomyopathy. *The American Journal of Human Genetics*, 88(5):635–642.

Hildebrand, F., Meyer, A., and Eyre-Walker, A. (2010). Evidence of selection upon genomic gc-content in bacteria. *PLoS Genet*, 6(9):e1001107.

Hulgan, T., Kallianpur, A. R., Guo, Y., Barnholtz-Sloan, J. S., Gittleman, H., Brown, T. T., Ellis, R., Letendre, S., Heaton, R. K., Samuels, D. C., et al. (2019). Peripheral blood mitochondrial dna copy number obtained from genome-wide genotype data is associated with neurocognitive impairment in persons with chronic hiv infection. *Journal of acquired immune deficiency syndromes (1999)*, 80(4):e95.

Hurtado-Roca, Y., Ledesma, M., Gonzalez-Lazaro, M., Moreno-Loshuertos, R., Fernandez-Silva, P., Enriquez, J. A., and Laclaustra, M. (2016). Adjusting mtdna quantification in whole blood for peripheral blood platelet and leukocyte counts. *PLoS One*, 11(10):e0163770.

Longchamps, R. J., Castellani, C. A., Newcomb, C. E., Sumpter, J. A., Lane, J., Grove, M. L., Guallar, E., Pankratz, N., Taylor, K. D., Rotter, J. I., et al. (2019). Evaluation of mitochondrial dna copy number estimation techniques. *bioRxiv*, page 610238.

Longchamps, R. J., Castellani, C. A., Yang, S. Y., Newcomb, C. E., Sumpter, J. A., Lane, J., Grove, M. L., Guallar, E., Pankratz, N., Taylor, K. D., et al. (2020). Evaluation of mitochondrial dna copy number estimation techniques. *PloS one*, 15(1):e0228166.

Lowell, B. B. and Shulman, G. I. (2005). Mitochondrial dysfunction and type 2 diabetes. *Science*, 307(5708):384–387.

Malik, A. N. and Czajka, A. (2013). Is mitochondrial dna content a potential biomarker of mitochondrial dysfunction? *Mitochondrion*, 13(5):481–492.

McBride, H. M., Neuspiel, M., and Wasiak, S. (2006). Mitochondria: more than just a powerhouse. *Current biology*, 16(14):R551–R560.

Nunnari, J. and Suomalainen, A. (2012). Mitochondria: in sickness and in health. *Cell*, 148(6):1145–1159.

Nygaard, V., Rødland, E. A., and Hovig, E. (2016). Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics*, 17(1):29–39.

Pankratz, N. (2018). Mitopipeline: Generating mitochondrial copy number estimates from snp array data in genvisis.

peng Yan, H., Li, M., lan Lu, X., min Zhu, Y., Ou-Yang, W.-X., hui Xiao, Z., Qiu, J., and jie Li, S. (2018). Use of plasma mitochondrial dna levels for determining disease severity and prognosis in pediatric sepsis: a case control study. *BMC pediatrics*, 18(1):267.

Pietiläinen, K. H., Naukkarinen, J., Rissanen, A., Saharinen, J., Ellonen, P., Keränen, H., Suomalainen, A., Götz, A., Suortti, T., Yki-Järvinen, H., et al. (2008). Global transcript profiles of fat in monozygotic twins discordant for bmi: pathways behind acquired obesity. *PLoS medicine*, 5(3):e51.

Podlesniy, P., Figueiro-Silva, J., Llado, A., Antonell, A., Sanchez-Valle, R., Alcolea, D., Lleo, A., Molinuevo, J. L., Serra, N., and Trullas, R. (2013). Low cerebrospinal fluid concentration of mitochondrial dna in preclinical alzheimer disease. *Annals of neurology*, 74(5):655–668.

Pyle, A., Anugrha, H., Kurzawa-Akanbi, M., Yarnall, A., Burn, D., and Hudson, G. (2016). Reduced mitochondrial dna copy number is a biomarker of parkinson's disease. *Neurobiology of Aging*, 38:216–e7.

Pyle, A., Brennan, R., Kurzawa-Akanbi, M., Yarnall, A., Thouin, A., Mollenhauer, B., Burn, D., Chinnery, P. F., and Hudson, G. (2015). Reduced cerebrospinal fluid mitochondrial dna is a biomarker for early-stage parkinson's disease. *Annals of neurology*, 78(6):1000–1004.

Qian, Y., Butler, T. J., Opsahl-Ong, K., Giroux, N. S., Sidore, C., Nagaraja, R., Cucca, F., Ferrucci, L., Abecasis, G. R., Schlessinger, D., et al. (2017). fastmitocalc: an ultra-fast program to estimate mitochondrial dna copy number from whole-genome sequences. *Bioinformatics*, 33(9):1399–1401.

Reznik, E., Miller, M. L., Şenbabaoğlu, Y., Riaz, N., Sarungbam, J., Tickoo, S. K., Al-Ahmadie, H. A., Lee, W., Seshan, V. E., Hakimi, A. A., et al. (2016). Mitochondrial dna copy number variation across human cancers. *elife*, 5:e10769.

Romiguier, J. and Roux, C. (2017). Analytical biases associated with gc-content in molecular evolution. *Frontiers in Genetics*, 8:16.

Scatena, R. (2012). Mitochondria and cancer: a growing role in apoptosis, cancer cell metabolism and dedifferentiation. In *Advances in Mitochondrial Medicine*, pages 287–308. Springer.

Shen, M., Zhang, L., Bonner, M. R., Liu, C.-S., Li, G., Vermeulen, R., Dosemeci, M., Yin, S., and Lan, Q. (2008). Association between mitochondrial dna copy number, blood cell counts, and occupational benzene exposure. *Environmental and molecular mutagenesis*, 49(6):453–457.

Stock, D., Leslie, A. G., and Walker, J. E. (1999). Molecular architecture of the rotary motor in atp synthase. *Science*, 286(5445):1700–1705.

Tin, A., Grams, M. E., Ashar, F. N., Lane, J. A., Rosenberg, A. Z., Grove, M. L., Boerwinkle, E., Selvin, E., Coresh, J., Pankratz, N., et al. (2016). Association between mitochondrial dna copy number in peripheral blood and incident ckd in the atherosclerosis risk in communities study. *Journal of the American Society of Nephrology*, 27(8):2467–2473.

van Heesch, S., Mokry, M., Boskova, V., Junker, W., Mehon, R., Toonen, P., de Bruijn, E., Shull, J. D., Aitman, T. J., Cuppen, E., et al. (2013). Systematic biases in dna copy number originate from isolation procedures. *Genome biology*, 14(4):R33.

Wang, L., Xie, L., Zhang, Q., Cai, X., Tang, Y., Wang, L., Hang, T., Liu, J., and Gong, J. (2015). Plasma nuclear and mitochondrial dna levels in acute myocardial infarction patients. *Coronary artery disease*, 26(4):296.

Wei, W., Keogh, M. J., Wilson, I., Coxhead, J., Ryan, S., Rollinson, S., Griffin, H., Kurzawa-Akanbi, M., Santibanez-Koref, M., Talbot, K., et al. (2017). Mitochondrial dna point mutations and relative copy number in 1363 disease and control human brains. *Acta neuropathologica communications*, 5(1):13.

Weijers, R. N. and Bekedam, D. J. (2007). Relationship between gestational diabetes mellitus and type 2 diabetes: evidence of mitochondrial dysfunction. *Clinical chemistry*, 53(3):377–383.

Westermann, B. (2010). Mitochondrial fusion and fission in cell life and death. *Nature reviews Molecular cell biology*, 11(12):872–884.

Whitaker, R. M., Stallons, L. J., Kneff, J. E., Alge, J. L., Harmon, J. L., Rahn, J. J., Arthur, J. M., Beeson, C. C., Chan, S. L., and Schnellmann, R. G. (2015). Urinary mitochondrial dna is a biomarker of mitochondrial disruption and renal dysfunction in acute kidney injury. *Kidney international*, 88(6):1336–1344.

Wong, J., McLennan, S., Molyneaux, L., Min, D., Twigg, S., and Yue, D. (2009). Mitochondrial dna content in peripheral blood monocytes: relationship with age of diabetes onsetand diabetic complications. *Diabetologia*, 52(9):1953.

Zhang, P., Lehmann, B. D., Samuels, D. C., Zhao, S., Zhao, Y.-Y., Shyr, Y., and Guo, Y. (2017a). Estimating relative mitochondrial dna copy number using high throughput sequencing data. *Genomics*, 109(5-6):457–462.

Zhang, Y., Guallar, E., Ashar, F. N., Longchamps, R. J., Castellani, C. A., Lane, J., Grove, M. L., Coresh, J., Sotoodehnia, N., Ilkhanoff, L., et al. (2017b). Association between mitochondrial dna copy number and sudden cardiac death: findings from the atherosclerosis risk in communities study (aric). *European heart journal*, 38(46):3443–3448.